OmniMotion-X: Versatile Multimodal Whole-Body Motion Generation

Guowei Xu[1*], Yuxuan Bian[2*], Ailing Zeng[5†], Zhuo Chen[1], Mingyi Shi[3], Shaoli Huang[4], Wen Li[1†], Lixin Duan[1], Qiang Xu[2]
[1]UESTC [2]CUHK [3]HKU [4]Tencent [5]Independent Researcher
{xuguowei368, yuxuanbian23, ailingzengzzz, liwenbnu, lxduan}@gmail.com

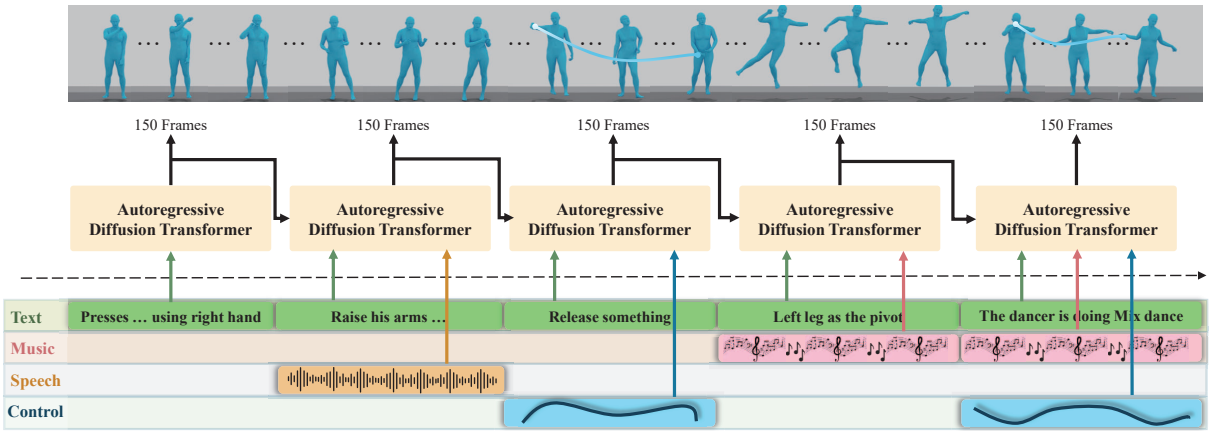


Figure 1. We present *OmniMotion-X*, a unified sequence-to-sequence autoregressive motion diffusion transformer designed for flexible and interactive whole-body human motion generation. It supports a variety of tasks, including text-to-motion, music-to-dance, speech-to-gesture, and globally spatial-temporal controllable motion generation, which encompasses motion prediction, in-betweening, completion, and joint/trajectory-guided synthesis. These conditions can be combined in various ways to enable versatile motion generation.

Abstract

*This paper introduces **OmniMotion-X**, a versatile multimodal framework for whole-body human motion generation, leveraging an autoregressive diffusion transformer in a unified sequence-to-sequence manner. OmniMotion-X efficiently supports diverse multimodal tasks, including text-to-motion, music-to-dance, speech-to-gesture, and global spatial-temporal control scenarios (e.g., motion prediction, in-betweening, completion, and joint/trajectory-guided synthesis)—as well as flexible combinations of these tasks. Specifically, we propose the use of reference motion as a novel conditioning signal, substantially enhancing the consistency of generated content, style, and temporal dynamics crucial for realistic animations. To handle multimodal conflicts, we introduce a progressive weak-to-strong mixed conditions training strategy. To enable high-quality multimodal training, we construct **OmniMoCap-X**, the largest unified multimodal motion dataset to date, integrating 28 publicly available MoCap sources across 10 distinct tasks, standardized to the SMPL-X format at 30 fps. To ensure detailed and consistent annotations, we render sequences into videos and use GPT-4o to automatically generate structured and hierarchical captions, capturing both low-level actions and high-level semantics. Extensive experimental evaluations confirm that OmniMotion-X significantly surpasses existing methods, demonstrating state-of-the-art performance across multiple multimodal tasks and enabling the interactive generation of realistic, coherent, and controllable long-duration motions. The code and dataset are publicly available at https://github.com/GuoweiXu368/OmniMocap-X.*

1. Introduction

Multimodal whole-body human motion generation plays critical roles in animation [11], gaming [44], virtual reality [23], and embodied intelligence [56] across diverse input conditions, including text, audio, and trajectory, etc. The limited multimodal data and task-specific model designs prevent existing motion generation methods from supporting multimodal whole-body human motion generation.

*Equal contribution; †Corresponding authors.

Model Type	Method	Tasks					Reference Motion	Mixed-condition	Whole-Body	Training Data	
		T2M	M2D	S2G	GSTC(S)	GSTC(D)				Datasets	Hours
DiT	MDM [66]	✓	×	×	×	×	×	×	×	2	28.6
DiT	MCM [47]	✓	✓	✓	×	×	×	×	×	3	109.8
DiT	LMM [85]	✓	✓	✓	×	×	×	✓	✓	16	-
DiT	MotionCraft [7]	✓	✓	✓	×	×	×	×	✓	3	48.4
AR	MoMask [22]	✓	×	×	×	×	×	×	×	2	28.6
AR	MotionGPT [30]	✓	×	×	×	×	×	×	×	2	28.6
AR	M^3GPT [55]	✓	✓	×	×	×	×	✓	×	3	164
AR-DiT	AMD [24]	✓	✓	×	×	×	✓	×	×	4	85.87
AR-DiT	DART [88]	✓	×	×	✓	✓	✓	✓	×	2	43.5
AR-DiT	*OmniMotion-X* (Ours)	✓	✓	✓	✓	✓	✓	✓	✓	**28**	**286.2**

Table 1. Comparison between *OmniMotion-X* and existing human motion generation methods. "GSTC(S)" and "GSTC(D)" denote global spatial-temporal controllable motion generation, where "S" and "D" indicate sparse and dense controlled joints, respectively. "Reference Motion" originates from user-designed or previously generated motion. "Mixed-condition" refers to the simultaneous occurrence of multiple conditions during training. "Datasets" indicates the total number of datasets used for training, while "Hours" represents the longest training dataset duration. For methods like MoMask, trained separately on HumanML3D or KIT, the duration of the HumanML3D dataset is considered.

Due to the high cost of mocap data collection and labeling, most datasets focus on single domains such as Text-to-Motion (T2M) [21, 61], Music-to-Dance (M2D) [39, 40], Speech-to-Gesture (S2G) [50, 80], Human-Object Interaction (HOI) [6, 17, 27], Human-Scene Interaction (HSI) [32], and Human-Human Interaction (HHI) [43, 75], each with inconsistent data formats (*e.g.*, BVH, SMPL-(H/X) and 3D Keypoints) and control conditions (*e.g.*, Text captions, Audio, and Trajectories). To tackle motion generation across diverse scenarios, it is crucial to build a unified framework that utilizes large-scale, diverse data to achieve more generalized representations.

Although recent works [7, 24, 47, 48, 85] attempt to unify multitasks in one model, they meet challenges in multimodal control modeling, versatile tasks, and high-quality motion generation (as shown in Tab. 1): (1) **Independent Model Training.** Previous approaches train separate models for each modality, limiting simultaneous control across inputs [24]. (2) **Additional Control Branches.** Some methods add separate control branches for each condition, limiting interaction between them [7, 47]. (3) **Conflict Granularity Training.** Existing methods use mixed training by combining high-level semantic conditions with low-level controls, which hampers effective control at different levels and leads to optimization challenges [48, 55, 85]. Similar phenomena have also been observed in video generation [4, 45]. In addition to modeling challenges, relevant works also introduce large-scale motion datasets from multiple tasks and modalities [42, 46, 48, 54, 85], they still exhibit the following significant shortcomings (see comparisons in Tab. 2): (1) **Low Motion Quality.** Datasets expanded with non-mocap motion estimation exhibit "garbage in, garbage out" effects, leading to poor-quality motion [46, 54, 85]. (2) **Text Inconsistency.** Inconsistent text annotations or expanded LLM descriptions lead to uneven text quality and hallucination issues [42, 48, 54]. and (3) **limited tasks.** They focus on common tasks, limiting applicability in diverse ones like HOI, HSI, and HHI [7, 54, 85]. These limitations hinder the development of a unified, high-quality dataset for multimodal whole-body human motion generation across diverse scenarios.

To address existing challenges, we propose *OmniMotion-X*, a unified framework for multimodal whole-body human motion generation, and *OmniMoCap-X*, the largest unified multimodal mocap motion dataset. Specifically, *OmniMotion-X* employs Diffusion Transformers (DiTs) [60], incorporates multimodal conditions by concatenating condition tokens as prefix context, and adopts a progressive weak-to-strong mixed conditions training strategy to gradually constrain motion from high-level semantics to dense spatial-temporal alignment. Notably, unlike previous methods, *OmniMotion-X* introduces a novel generation paradigm that utilizes reference motion (*e.g.*, user-provided or model-predicted motion) as a special condition. This significantly enhances generated motion quality and achieves consistency between reference and generated motion, creating an effective clip-by-clip autoregressive motion diffusion. This enables *OmniMotion-X* to support autoregressive interactive generation with strong temporal alignment. Furthermore, *OmniMotion-X* unifies various Global Spatial-Temporal Controllable Generation tasks through spatial-temporal masking strategies. To guarantee high-quality motion generation training, we collect high-quality mocap datasets that support diverse motion generation tasks, unify them under the SMPL-X [59] format with standard world coordinate systems, and automatically generate hierarchical text captions by rendering motions into videos and annotating them with vision language models (VLMs). This dataset contains about 286.2 *hours* and integrates multimodal control conditions, supporting versatile tasks, including T2M, M2D, S2G, HOI,

Dataset	Tasks						Whole-Body	Motion Source Mocap / Total	Caption Source	Hierarchical Caption	Frames	Hours
	T2M	M2D	S2G	HOI	HSI	HHI						
Motion-X [46]	✓	×	×	×	×	×	✓	2 / 9	T (9)	✓	15.6M	144.2
OMG [42]	✓	×	×	×	×	✓	×	9 / 13	-	×	22.3M	206.5
MotionUnion [54]	✓	×	×	×	×	×	×	4 / 15	-	×	30M	260
MotionVerse [85]	✓	✓	✓	×	×	×	✓	5 / 16	-	×	**100M**	-
OmniMoCap-X (Ours)	✓	✓	✓	✓	✓	✓	✓	**21** / **28**	V + T (28)	✓	64.3M	**286.2**

Table 2. Comparisons between *OmniMoCap-X* and existing merged datasets. "Mocap Source" indicates the proportion of mocap datasets. "Caption Source" specifies the method for completing missing descriptions: "-" (no completion), "V" (visual information), and "T" (textual information). "Hierarchical Caption" shows if captions include hierarchical text.

HSI, and HHI.

In summary, our core contributions are as follows:

- We propose *OmniMotion-X*, a multimodal autoregressive diffusion transformer for versatile whole-body motion generation. By introducing reference motion, *OmniMotion-X* significantly enhances consistent content, style, and temporal dynamics generation.
- We propose a progressive weak-to-strong mixed conditions training strategy to effectively handle multi-granular constraints.
- We construct *OmniMoCap-X*, the largest unified multimodal mocap motion dataset with a unified SMPL-X format, and provides consistent, detailed, and structured text captions to serve versatile motion generation tasks.
- Extensive experiments show that *OmniMotion-X* achieves state-of-the-art performance across various tasks, including T2M, S2G, M2D, and GSTC. These evaluations were conducted on our more challenging test sets comprising 280 samples uniformly sampled from our *OmniMoCap-X* across diverse scenarios.

2. Related Work

2.1. Human Motion Generation

Existing methods are typically categorized based on input conditions into three main types: single-modal, cross-modal, and multimodal. Single-modal motion generation uses control conditions from the same modality as the motion, including motion prediction [10, 74], in-betweening [13], joint/trajectory-guided synthesis [14, 73, 88], and body-shape-conditioned motion generation [67, 76]. Cross-modal motion generation uses control conditions from different modalities, including text in T2M [12, 22, 30, 42, 53, 66, 82, 84, 86], music in M2D [35, 40, 65, 68], speech in S2G [8, 18, 50, 80], target object positions in HOI [17, 51, 77], scene layouts in HSI [26, 32, 33], and partner movements in HHI [43, 71, 75]. However, these approaches typically rely on task-specific architectures, significantly limiting their cross-task generalization capabilities and practical applications. Recently, researchers [7, 24, 38, 47, 55, 85] have begun exploring multimodal motion generation, with approaches falling into three primary categories:

- **Separate Model Training.** These approaches [24] typically train independent models for different conditional modalities, making multimodal motion generation fundamentally unattainable.
- **Additional Control Branch.** These approaches [7, 47] incorporate separate control branches into the backbone architecture, each dedicated to a specific condition, resulting in limited interaction between different conditions.
- **Unified Multimodal Condition Modeling.** These approaches [48, 78, 85, 87] use a single model to learn mappings from diverse modalities to target motions, enhancing multi-modal adaptability. However, challenges arise from different constraints across modalities: text provides semantic guidance, spatial-temporal controls impose physical constraints, while audio inputs enforce rhythmic alignment. These disparate constraints often result in compromised controllability and optimization difficulties.

2.2. Human Motion Dataset

Current human motion generation datasets are predominantly task-specific. Different tasks rely on separate collections: text-to-motion (T2M) uses datasets with text captions [21, 44, 46, 61, 62], music-to-dance (M2D) requires music-annotated datasets [9, 35, 39, 40], and speech-to-gesture (S2G) employs datasets with paired speech [18, 49, 50, 80]. Similarly, interaction datasets remain disconnected across human-object [6, 17, 27, 31, 37, 51, 81, 83], human-scene [32], and human-human interaction [43, 75]. This fragmentation constrains multimodal motion generation capabilities. Integrating these datasets presents challenges due to inconsistent motion formats. Datasets vary widely in their representations: some use SMPL [6, 16, 39, 43], others employ SMPL-X [18, 50, 80], while many rely on BVH [2, 3, 25, 57, 69] or keypoint representations [28, 29]. This diversity makes format standardization extremely difficult. Recent works [42, 46, 48, 54, 85] have attempted to integrate multiple datasets, increasing scale and incorporating some multimodal conditions. However, they still exhibit several limitations:

- **Low-quality Motions**: Most integrated datasets derive from non-mocap sources with significant estimation errors [16, 46, 48, 54, 85], creating a "garbage in, garbage

out" effect that compromises motion generation quality.

- **Text Inconsistency**: Existing datasets [42, 54, 85] use inconsistent text annotations (*e.g.*, action labels versus semantic descriptions) or rely on LLMs for text refinement without visual motion data, causing variable text quality and hallucinations [46, 48].
- **Limited Generation Scenarios**: Most datasets lack support for critical interaction tasks including HOI, HSI [42, 46, 48, 54, 85], and HHI [46, 54, 85], severely limiting their applicability to complex scenarios.

3. OmniMoCap-X Dataset

To address the data challenges—the scarcity of high-quality multitask motion datasets, inconsistent textual annotations and motion representations—we implement a three-pronged approach: (1) integrating diverse motion capture datasets with multitask support, (2) establishing a unified motion representation, and (3) developing high-quality motion captions derived from visual-textual annotations.

3.1. Mocap Dataset Composition

While existing approaches prioritize dataset scale over quality by incorporating substantial amounts of non-mocap data [46, 48, 54, 85], leading to degraded motion generation quality ("garbage in, garbage out"), our work emphasizes data quality. We systematically curate open-source mocap datasets and classify them into five acquisition categories: Marker with manual correction, Marker (Vicon), IMU, Multi-View RGB, and Single-View RGB (detailed in Tab. 3). Our dataset supports diverse motion generation tasks, including text-to-motion, motion-to-dance, speech-to-gesture, and various interactions (human-object, human-scene, and human-human), providing a high-quality multimodal foundation. Comprising 64.3 million frames and 286.2 hours of data, our approach achieves comprehensive task coverage while maintaining superior data quality. Visualizations of specific interaction scenarios are showcased in the supplementary material.

Task	Dataset	Frames	Hours	Mocap	Format
T2M	Mixamo [2]	0.4M	1.9	Marker-M	BVH
	KIT [61]	2.3M	6.4	Marker-V	SMPL
	OMOMO [37]	1.1M	2.5	Marker-V	SMPL-X
	IDEA400 [46]	1.7M	15.7	SV-RGB	SMPL-X
	100Style [57]	4.8M	22.1	IMU	BVH
	HumanML3D [21]	26.5M	65.7	Marker-V	SMPL
M2D	Choreomaster [9]	0.1M	1.2	Marker-M	FBX
	Finedance [40]	0.8M	7.7	Marker-V	SMPL-H
	Phantomdance [36]	1.0M	9.5	Marker-M	SMPL
	AIST++ [39]	1.1M	5.2	MV-RGB	SMPL
	Motorica [3]	2.7M	12.4	Marker-V	BVH
	AIOZ [35]	6.6M	60.8	SV-RGB	SMPL
S2G	BEAT2 [50]	6.9M	64.2	Marker-V	SMPL-X
HHI	Humansc3d [19]	0.2M	1.1	Marker-V	SMPL-X
	InterHuman [43]	4.5M	20.8	MV-RGB	SMPL
	Inter-x [75]	16.2M	37.5	Marker-V	SMPL-X
HOI	Arctic [17]	0.2M	2.0	Marker-V	SMPL-X
	TACO [51]	0.4M	3.3	Marker-V	MANO
	Fit3d [20]	0.4M	2.5	Marker-V	SMPL-X
	Behave [6]	0.4M	4.1	MV-RGB	SMPL-X
	Chairs [31]	1.0M	9.5	Marker-V	SMPL-X
	HOI-M3 [83]	2.3M	10.7	MV-RGB	SMPL
	Oaklnkv2 [81]	4.0M	36.8	Marker-V	SMPL-X
HSI	EMDB [33]	0.03M	0.3	IMU	SMPL
	Rich [26]	0.1M	0.8	MV-RGB	SMPL-X
	Lafan1 [25]	1.0M	4.5	Marker-V	BVH
	Trumans [32]	1.2M	11.2	Marker-V	SMPL-X
	Circle [5]	4.4M	10.1	Marker-V	SMPL-X
All	*OmniMoCap-X*	64.3M	286.2	Mixed	SMPL-X

Table 3. Composition of *OmniMoCap-X* dataset, unifying motion formats into SMPL-X with captions. We select 28 publicly available high-quality datasets across various tasks. Frames and Hours are computed based on raw dataset FPS. MoCap represents data capture methods, ranked by quality: Marker with manual correction (Marker-M), Vicon Marker (Marker-V), IMU, Multi-View RGB (MV-RGB), and Single-View RGB (SV-RGB). Format specifies the original motion format.

3.2. Unified Motion Representation

We standardize diverse motion formats (BVH [34], FBX [72], and SMPL (-H) [52]) into the SMPL-X [59] format to support our unified multimodal model. First, we convert these formats into the Motion-X [46] SMPL-X format, comprising root orientation, pose parameters (body, hand, and jaw), facial expressions, facial shape, translation, and body shape parameters. We then normalize the translation scale and initial root orientation across datasets to establish a consistent coordinate system. Additionally, we resample all datasets to 30 frames per second (FPS) to enhance the model's ability to capture temporal patterns. The specific conversion process and normalization for different datasets are provided in the supplementary material.

To enhance generalization ability, we extend the widely-used body-only representation [21] to a whole-body format. Specifically, the pose of the i-th frame is a tuple $\mathbf{p}_i = (\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, \mathbf{j}^p, \mathbf{j}^v, \mathbf{c}^f, \mathbf{f})$, where $\dot{r}^a \in \mathbb{R}$ is root angular velocity around the Y-axis; $\dot{r}^x, \dot{r}^z \in \mathbb{R}$ are root linear velocity in the XZ plane; $r^y \in \mathbb{R}$ is root height, $\mathbf{j}^p \in \mathbb{R}^{3N-1}$ are local joint positions; Notably, $\mathbf{j}^r \in \mathbb{R}^{6N'}$ are joint 6D rotations of SMPL-X; $\mathbf{j}^v \in \mathbb{R}^{3N}$ are joint velocities; $\mathbf{c}^f$ is binary features obtained by thresholding the heel and toe joint velocities to emphasize the foot ground contacts. Here, N refers to 127 whole-body joints extracted from SMPL-X [59], while N' represents 53 joints spanning the body, hands, and jaw. For facial representation, we adopt the Flame Format [41], encoding facial features as $\mathbf{f} \in \mathbb{R}^{100}$.

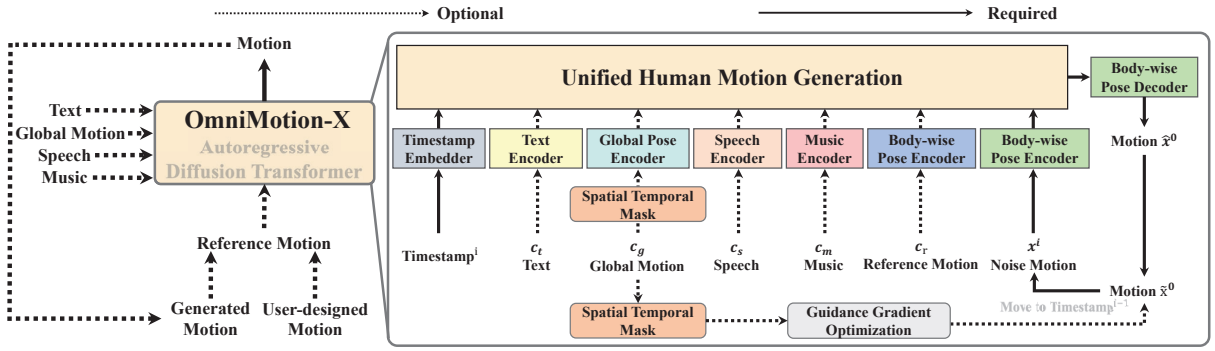


Figure 2. Overview of *OmniMotion-X*, a unified multimodal autoregressive transformer diffusion model for whole-body human motion generation. *OmniMotion-X* integrates text, global motion, speech, music, and reference motion as conditions through condition-specific encoders mapped into a unified space. The model fuses multimodal information to produce coherent motion, with spatial-temporal guidance ensuring consistent global motion characteristics.

3.3. Consistent Visual-textual Motion Captions

Current datasets employ three annotation approaches:

- **LLM Refinement** [46, 54] employs LLMs like GPT-4 [1] to enhance textual descriptions. However, since LLMs cannot directly perceive motion data and rely solely on existing text, they frequently introduce hallucinations and fail to capture precise action details.
- **Manual Annotation** [21, 48, 62] effectively eliminates hallucinations but proves costly and difficult to scale, resulting in limited annotations with simple descriptions.
- **Motion-to-Text Model-Based Annotation** [15, 79] generates captions directly from raw motion. While these methods effectively capture low-level kinematic details (*e.g.*, left elbow bends 45 degrees), they lack the understanding of high-level semantics and action context.

To overcome these limitations, we propose an automated approach that integrates both visual and textual information for comprehensive motion annotation. Our method renders motion into videos and combines them with existing textual annotations (*e.g.*, descriptions, action labels, and task categories) as input to the state-of-the-art vision-language models (VLMs), GPT-4o [1]. This multimodal approach generates structured, hierarchical, and precise motion captions that ensure both annotation quality and expression richness. The quality of the texts was ensured through iterative prompt optimization and a comparative evaluation of various VLMs, with specific examples and detailed caption statistics provided in the supplementary material.

4. Method

The overview of *OmniMotion-X* is illustrated in Fig. 2. We propose a unified multimodal, autoregressive transformer-based diffusion model for whole-body human motion generation across multiple tasks. The model takes textual description $\mathbf{c}_t$ for semantic guidance, global motion $\mathbf{c}_g$ for spatial-temporal control, and speech $\mathbf{c}_s$ and music conditions $\mathbf{c}_m$ to ensure rhythmic and stylistic coherence. Furthermore, it incorporates reference motion $\mathbf{c}_r$ as a motion prior, derived from either previously generated clips or user-designed motion, thus providing fine-grained details unavailable in other conditions. This reference input is optional and can be set to null, which is essential for generating the initial motion when no user-specified prior exists. In addition, the mixed-condition setting allows multiple conditions to occur simultaneously during training, enabling the model to handle complex scenarios with diverse inputs. In this section, we present our approach in two parts: a unified framework for multimodal modeling and a progressive weak-to-strong mixed conditions training strategy that ensures precise motion control.

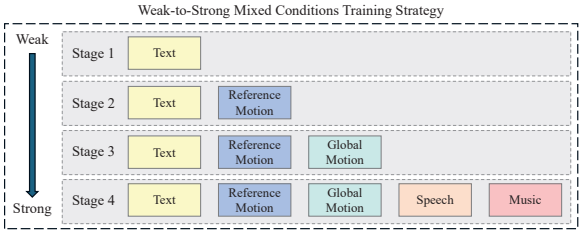


Figure 3. We propose the weak-to-strong conditions training strategy, establishing motion-semantic alignment with text, followed by progressive integration of stronger multimodal signals (reference motion, global motion, speech, music) for enhanced generation quality and controllability.

4.1. Unified Multimodal Modeling

We integrate multimodal conditions using Diffusion Transformers (DiTs) [60] by concatenating condition tokens as prefix contexts, allowing the model to process and fuse multimodal information efficiently.

Multimodal Conditions. Our framework integrates mul-

Method	R Precision↑			FID↓	Multimodal Dist.↓	Diversity→	MultiModality↑
	Top-1	Top-2	Top-3				
GT	$0.535^{\pm0.009}$	$0.725^{\pm0.009}$	$0.821^{\pm0.008}$	$0.013^{\pm0.005}$	$2.493^{\pm0.011}$	$9.194^{\pm0.093}$	-
MDM [66]	$0.063^{\pm0.004}$	$0.121^{\pm0.003}$	$0.169^{\pm0.003}$	$72.928^{\pm0.361}$	$8.376^{\pm0.018}$	$1.981^{\pm0.003}$	$0.394^{\pm0.010}$
MLD [12]	$0.084^{\pm0.002}$	$0.152^{\pm0.002}$	$0.209^{\pm0.003}$	$70.082^{\pm0.468}$	$8.281^{\pm0.025}$	$1.998^{\pm0.003}$	$0.439^{\pm0.022}$
MoMask [22]	$0.104^{\pm0.003}$	$0.163^{\pm0.003}$	$0.199^{\pm0.005}$	$69.361^{\pm0.365}$	$8.341^{\pm0.022}$	$2.123^{\pm0.002}$	$0.482^{\pm0.008}$
MoMask* [22]	$0.267^{\pm0.004}$	$0.414^{\pm0.004}$	$0.530^{\pm0.004}$	$17.428^{\pm0.400}$	$5.661^{\pm0.024}$	$6.772^{\pm0.013}$	$0.811^{\pm0.067}$
MotionCraft [7]	$0.176^{\pm0.004}$	$0.259^{\pm0.003}$	$0.319^{\pm0.004}$	$63.049^{\pm0.470}$	$7.936^{\pm0.021}$	$2.325^{\pm0.013}$	$0.557^{\pm0.039}$
MotionCraft* [7]	$0.236^{\pm0.004}$	$0.370^{\pm0.004}$	$0.489^{\pm0.004}$	$47.428^{\pm0.400}$	$7.424^{\pm0.024}$	$2.820^{\pm0.013}$	$0.863^{\pm0.067}$
OmniMotion-X (Ours)	$\underline{0.303}^{\pm0.004}$	$\underline{0.464}^{\pm0.006}$	$\underline{0.571}^{\pm0.005}$	$\underline{5.040}^{\pm0.293}$	$\underline{4.678}^{\pm0.019}$	$\mathbf{8.650^{\pm0.095}}$	$\mathbf{1.696^{\pm0.366}}$
OmniMotion-X (Ours + RM)	$\mathbf{0.346^{\pm0.005}}$	$\mathbf{0.511^{\pm0.007}}$	$\mathbf{0.629^{\pm0.006}}$	$\mathbf{3.199^{\pm0.293}}$	$\mathbf{4.106^{\pm0.019}}$	$\underline{8.009}^{\pm0.095}$	$\underline{1.143}^{\pm0.366}$

Table 4. Quantitative results of text-to-motion on the *OmniMoCap-X* test set. ↑ (↓) indicates that a larger (smaller) value is better. → indicates that a value closer to the GT is better. Bold entries indicate the best results, and underlined entries indicate the second-best results. All evaluations are repeated 20 times, reporting the mean and 95% confidence interval.

tiple condition modalities: text $\mathbf{c}_t$ providing semantic guidance; global motion $\mathbf{c}_g$ ensuring spatial-temporal consistency; speech $\mathbf{c}_s$ synchronizing gestures and lip movements with rhythm; music $\mathbf{c}_m$ supplying beat and style information for dance; and reference motion $\mathbf{c}_r$ serving as a motion prior. Notably, this reference motion condition—overlooked in previous multimodal motion generation [7, 47, 85]—enables our model to maintain precise spatial-temporal pattern consistency with the reference, substantially enhancing motion quality and coherence. Crucially, a reference motion provides rich spatio-temporal context that a single starting pose cannot. By typically spanning a duration similar to the target sequence, it captures fine-grained details and style, which are absent in a momentary posture. To fully leverage each modality and facilitate interaction between different conditions, we employ modality-specific encoders (*e.g.*, T5-XXL [64] for text, a wav encoder [50] for speech, Librosa [58] for music and body-wise encoding [7] for motion) to extract features from each modality. These features are aligned to match motion embedding dimensions using learnable linear projections, allowing us to concatenate all condition tokens as prefix context with noisy motion tokens during processing.

Multimodal Modeling. Overall, the unified multimodal modeling is formulated as follows:

$$c = [h_t(f_t(c_t)), h_g(f_g(c_g)), h_s(f_s(c_s)), h_m(f_m(c_m)), h_r(f_r(c_r))], \quad (1)$$

where f and h denote the modality-specific encoder and projection layer, respectively. The concatenated representation c is then fed into our DiT backbone as a conditioning prefix context, where attention mechanisms are employed to learn the correspondences among the various modalities. To constrain the physical properties of motion, we follow previous works [11, 66] by directly predicting motion $\hat{x}_0$ rather than noise. Consequently, our diffusion objective is defined as follows:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0 \sim q(x_0|c), t\sim[1,T]} \left[\|x_0 - G(x_t, t, c)\|_2^2 \right], \quad (2)$$

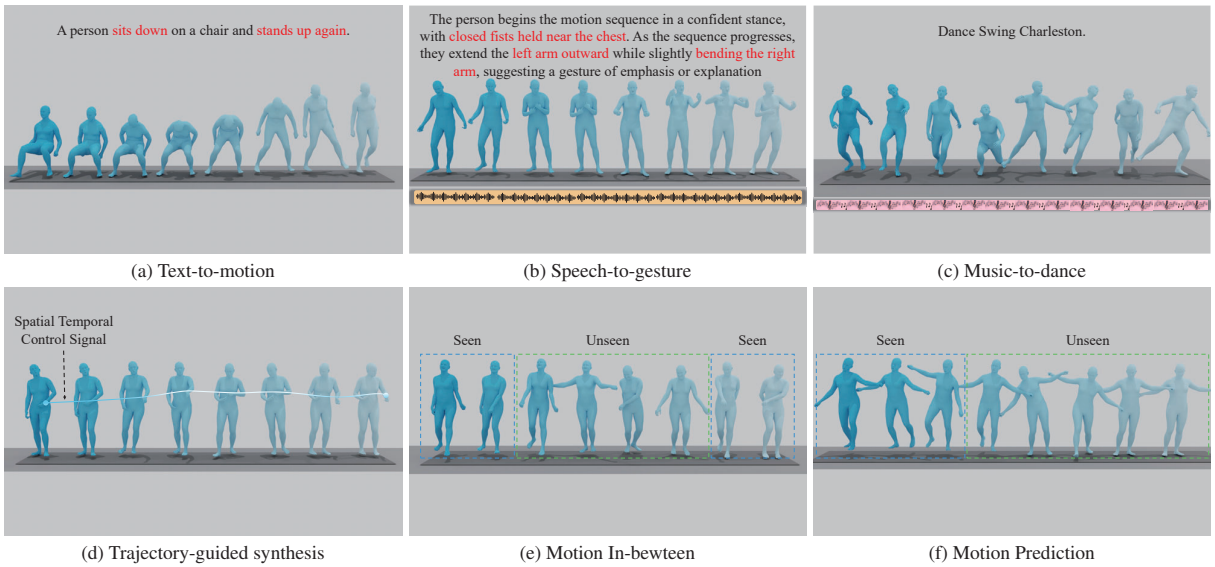
where $q(x_0|c)$ represents the data distribution, T is the maximum diffusion step. G denotes the learned denoising function, while x_t represents the noisy motion at step t, expressed as $x_t = [p_0^t, p_1^t, \ldots, \mathbf{p}_N^t]$, where each $\mathbf{p}_i^t$ corresponds to the i_{th} pose in motion at step t. Further details are available in the supplementary material.

Method	Text-motion quality			Geometric control			
	FID↓	R Prec.↑	Multi.↓	Ctrl L2↓	Skating↓	Fail@20↓	Fail@50↓
GT	0.013	0.821	2.493	-	-	-	-
OmniControl	<u>63.725</u>	<u>0.392</u>	<u>8.011</u>	<u>0.820</u>	<u>0.089</u>	<u>0.844</u>	<u>0.794</u>
OmniMotion-X	**4.224**	**0.682**	**4.377**	**0.424**	**0.004**	**0.516**	**0.330**

Table 5. Quantitative results of global spatiotemporal controllable generation on the *OmniMoCap-X* test set. We report text-motion quality metrics and geometric control metrics in a unified table. Bold entries indicate the best results, and underlined entries indicate the second-best results.

4.2. Weak-to-Strong Progressive Training Strategy

Considering our diverse multi-granularity conditioning inputs, we empirically observed that employing all conditions within a single-stage training paradigm leads to difficulties in directly learning the correlation between motion and conditions. Moreover, the model tends to overfit strongly constrained low-level control signals such as fine-grained reference motion and spatiotemporal joint control. While these signals appear dominant, they can suppress other modalities such as text, ultimately compromising overall controllability. To address this, we implement weak-to-strong mixed conditions training strategy: as shown in Fig. 3 initial text-conditioned learning establishes motion-semantic alignment, followed by progressive integration of finer-grained conditions, including reference motion, global motion, and audio signals. This progressive approach enables the model to effectively adapt to different conditions, ensuring high-quality and flexible motion generation while precisely adhering to multimodal conditions.



(a) Text-to-motion (b) Speech-to-gesture (c) Music-to-dance

(d) Trajectory-guided synthesis (e) Motion In-bewteen (f) Motion Prediction

Figure 4. **Diverse motion synthesis capabilities of *OmniMotion-X*.** *OmniMotion-X* supports multiple tasks: (a) text-to-motion, (b) speech-to-gesture, (c) music-to-dance, (d) trajectory-guided motion, (e) motion in-betweening, and (f) motion prediction.

Method	Speech-to-Gesture				Music-to-Dance		
	FID (Whole-Body)↓	FID (Hands)↓	$Face_{MSE}\downarrow$	Diversity↑	FID (Whole Body)↓	FID (Hands)↓	Diversity↑
MotionCraft	<u>3.422</u>	**5.370**	<u>0.182</u>	<u>1.003</u>	**9.875**	<u>7.099</u>	<u>3.798</u>
OmniMotion-X(Ours)	**2.641**	<u>9.095</u>	**0.045**	**1.664**	<u>16.209</u>	**5.827**	**4.716**

Table 6. **Results of S2G in BEAT2 and M2D in AIST++, FineDance and Phantomdance.** We respectively evaluate the $FID_{WholeBody}$, FID_{Hands}, $Face_{MSE}$, and diversity for S2G and the $FID_{WholeBody}$, FID_{Hands}, and diversity for M2D. Bold entries indicate the best results, and underlined entries indicate the second-best results.

5. Experiments

5.1. Implementation Details

Our method employs a Transformer Encoder architecture [70] with 8 layers and 8 attention heads, featuring a hidden dimension of $d_{\text{model}} = 1536$ (128×12), where 128 represents the embedding size per each of the 12 body parts, and a feedforward dimension of 3072. Training is performed on a single H800 GPU through progressive conditioning: starting with text-only training for $460K$ steps, then adding reference motion for another $460K$ steps, followed by global spatiotemporal control for $230K$ steps, and finally incorporating full audio conditions for $920K$ steps. The corresponding batch sizes are 48, 48, 48, and 16, respectively. We optimize with AdamW using an initial learning rate of 1×10^{-4} (reset for new conditions) and a cosine schedule decaying to 1×10^{-5} within the first $460K$ steps. The default length of motion reference and prediction is 150. During training and testing, the reference motion $\mathbf{c}_r$ is provided by the preceding segment of the ground-truth sequence, which shares the same temporal length as the motion to be generated, or is set to null. At inference, this capability is extended to accommodate user-defined motions, enabling customized and interactive generation. More implementation details are provided in supplementary material.

5.2. Quantitative Results

We evaluate *OmniMotion-X* on four tasks: Text-to-Motion (T2M), Global Spatiotemporal Controllable Generation (GSTC), Music-to-Dance (M2D), and Speech-to-Gesture (S2G). For T2M and GSTC evaluations, test sets are uniformly sampled across all datasets (10 samples from each). The M2D benchmark integrates test sequences from AIST++ [39], FineDance [40], and PhantomDance [36], with S2G evaluation conducted on BEAT2 [50].

T2M. Tab. 4 shows that previous methods, constrained by small-scale datasets, struggle to generalize in generating diverse and complex motions. In contrast, *OmniMotion-X* demonstrates a significant advantage, outperforming not only baseline methods trained on small datasets (MDM [66], MLD [12], MoMask [22], MotionCraft [7]) but also MoMask* and MotionCraft*, which are trained on our *OmniMoCap-X*. Compared to MoMask* and MotionCraft*,

Task	Method	FID↓	R Precision↑ Top-3	Multimodal Dist.↓	Diversity →
T2M	w/o TrSt	9.574	0.232	6.853	3.118
	Ours	**5.040**	**0.571**	**4.678**	**8.650**
GSTC	w/o TrSt	10.247	0.491	6.130	2.438
	Ours	**4.224**	**0.686**	**4.377**	**6.292**

Table 7. Ablation study on the weak-to-strong training strategy. Bold entries indicate the best results, and underlined entries indicate the second-best results.

OmniMotion-X uses stronger text encoder (T5-XXL [64]), which models complex texts more effectively than CLIP [63]. In addition, it adopts a unified multimodal framework with reference motion during training, enabling the model to learn more detailed whole-body motion representations, leading to better generation quality and generalization. We observed a significant performance enhancement when conditioning on the preceding ground-truth motion at test time, which validates the reference motion's role as a novel signal for preserving consistency in content, style, and temporal dynamics.

GSTC. We adopt the cross-joint setup from OmniControl [73], simulating spatially dense control by controlling all joints. As shown in Tab. 5, due to the small dataset size, OmniControl [73] struggles to generalize in GSTC task. In contrast, our method demonstrates clear advantages, effectively following spatially dense control signals.

M2D and S2G. We directly compare our method with the MotionCraft [7] on the S2G and M2D tasks. As shown in Tab. 6, *OmniMotion-X* achieves competitive performance. The slightly higher FID is primarily because the test sets for the S2G and M2D tasks are relatively small and have a limited distribution. In contrast, *OmniMotion-X* is trained on *OmniMoCap-X*, which comprises data from diverse tasks including T2M, M2D, S2G, HHI, HOI, and HSI. This diverse training endows our method with enhanced generative diversity, causing a distributional gap with the test sets of the S2G and M2D tasks.

5.3. Ablation Study

The ablation study in Tab. 7 reveals two key findings: (1) Removing the progressive weak-to-strong conditioning training strategy compromises text-conditioned alignment, indicating that finer-grained controls can override coarse semantic constraints; (2) Mixed-condition training strategy impairs spatiotemporal control by introducing optimization conflicts due to physical constraints, which underscores the necessity of adopting a weak-to-strong conditioning strategy. Furthermore, our progressive training strategy not only resolves these conflicts but also significantly reduces computational overhead in the initial stages. This is attributed to its design, where multimodal conditions are introduced incrementally as training progresses, resulting in a lower computational load in the early phases. In contrast, a conventional end-to-end approach that jointly trains on all conditions from the outset converges more slowly and faces greater challenges in achieving convergence.

5.4. Qualitative Results

As shown in Fig. 4, *OmniMotion-X* handles various multimodal generation, including T2M, M2D, S2G, and GSTC (covering motion prediction, in-betweening, and joint guidance). When combined with reference motion, the model generates conditioned motions that align with the reference. More visualizations are in the supplementary video.

6. Conclusion and Discussion

This work introduces the *OmniMotion-X* and the *OmniMoCap-X*. *OmniMotion-X* is an autoregressive diffusion model that integrates reference motion and multimodal conditions for precise whole-body motion control. It employs a progressive weak-to-strong mixed conditions strategy to effectively handle multi-granular constraints. *OmniMoCap-X* is the largest multimodal MoCap dataset, comprising 286.2 hours from 28 motion capture datasets, unified in SMPL-X with structured and consistent captions across 10 tasks. Experiments demonstrate that *OmniMotion-X* outperforms baselines, laying a strong foundation for large-scale multimodal motion generation.

Limitation and Future Work. While our method demonstrates versatility across multimodal conditions, it has key limitations. A primary limitation is the absence of explicit conditioning for physical interactions with scenes, objects, or other persons. Consequently, the model struggles to generate plausible interactive motions in complex environments. Future work should thus focus on incorporating interaction-aware conditions, such as conditioning on scene geometry and enforcing explicit physical constraints, to enhance realism and versatility. Furthermore, inference speed remains a significant concern. Our model (27.22M parameters) requires 2.14 seconds per sample, which is slower than text-conditioned diffusion models of a comparable scale. Notably, this inference time only increases to 3.00s when scaling the model tenfold to 355.13M parameters. This indicates the primary bottleneck is not the model's parameters but the extensive context length arising from aggregating diverse multimodal conditions. Future research should therefore explore more efficient architectures for handling these long-sequence multimodal interactions to accelerate inference.

Acknowledgement.

This work was supported by the Science, Technology and Innovation Project of Shenzhen Longhua District (No. 20260309G23410662).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5

[2] Adobe. Mixamo. https://www.mixamo.com. 3, 4

[3] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph.*, 42(4):44:1–44:20, 2023. 3, 4

[4] Tenglong Ao. Body of her: A preliminary study on end-to-end humanoid agent. *arXiv preprint arXiv:2408.02879*, 2024. 2

[5] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 4

[6] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2, 3, 4

[7] Yuxuan Bian, Ailing Zeng, Xuan Ju, Xian Liu, Zhaoyang Zhang, Wei Liu, and Qiang Xu. Motioncraft: Crafting whole-body motion with plug-and-play multimodal controls. *arXiv preprint arXiv:2407.21136*, 2024. 2, 3, 6, 7, 8

[8] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7352–7361, 2024. 3

[9] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3, 4

[10] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *ICCV*, pages 9544–9555, 2023. 3

[11] Rui Chen, Mingyi Shi, Shaoli Huang, Ping Tan, Taku Komura, and Xuelin Chen. Taming diffusion probabilistic models for character control. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 1, 6

[12] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *CVPR*, pages 18000–18010, 2023. 3, 6, 7

[13] Setareh Cohan, Guy Tevet, Daniele Reda, Xue Bin Peng, and Michiel van de Panne. Flexible motion in-betweening with diffusion models. In *ACM SIGGRAPH*, pages 1–9, 2024. 3

[14] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. *ECCV*, 2024. 3

[15] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022. 5

[16] Ke Fan, Shunlin Lu, Minyue Dai, Runyi Yu, Lixing Xiao, Zhiyang Dou, Junting Dong, Lizhuang Ma, and Jingbo Wang. Go to zero: Towards zero-shot motion generation with million-scale data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13336–13348, 2025. 3

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2, 3, 4

[18] Andrew Feng, Samuel Shin, and Youngwoo Yoon. A tool for extracting 3d avatar-ready gesture animations from monocular videos. In *Proceedings of the 15th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pages 1–7, 2022. 3

[19] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Learning complex 3d human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1343–1351, 2021. 4

[20] Mihai Fieraru, Mihai Zanfir, Silviu-Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[21] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, pages 5152–5161, 2022. 2, 3, 4, 5

[22] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *CVPR*, pages 1900–1910, 2024. 2, 3, 6, 7

[23] Xinying Guo, Mingyuan Zhang, Haozhe Xie, Chenyang Gu, and Ziwei Liu. Crowdmogen: Zero-shot text-driven collective motion generation. *arXiv preprint arXiv:2407.06188*, 2024. 1

[24] Bo Han, Hao Peng, Minjing Dong, Yi Ren, Yixuan Shen, and Chang Xu. Amd: Autoregressive motion diffusion. In *AAAI*, pages 2022–2030, 2024. 2, 3

[25] Félix G. Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. 39(4), 2020. 3, 4

[26] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 3, 4

[27] Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction. In *DAGM German Conference on Pattern Recognition*, pages 281–299. Springer, 2022. 2, 3

[28] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 3

[29] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale rgb-d database for arbitrary-view human action recognition. In *ACM MM*, pages 1510–1518, 2018. 3

[30] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *NeurIPS*, 2024. 2, 3

[31] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. *ICCV*, 3, 2022. 3, 4

[32] Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1737–1747, 2024. 2, 3, 4

[33] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. In *International Conference on Computer Vision (ICCV)*, 2023. 3, 4

[34] Christian Lauterbach, Michael Garland, Shubhabrata Sengupta, David Luebke, and Dinesh Manocha. Fast bvh construction on gpus. In *Computer Graphics Forum*, pages 375–384. Wiley Online Library, 2009. 4

[35] Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8673–8682, 2023. 3, 4

[36] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1279, 2022. 4, 7

[37] Jiaman Li, Jiajun Wu, and C Karen Liu. Object motion guided human motion synthesis. *ACM TOG*, 42(6):1–11, 2023. 3, 4

[38] Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. Genmo: A generalist model for human motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11766–11776, 2025. 3

[39] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with aist++: Music conditioned 3d dance generation, 2021. 2, 3, 4, 7

[40] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *ICCV*, pages 10234–10243, 2023. 2, 3, 4, 7

[41] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 4

[42] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–493, 2024. 2, 3, 4

[43] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *IJCV*, pages 1–21, 2024. 2, 3, 4

[44] Yihao Liao, Yiyu Fu, Ziming Cheng, and Jiangfeiyang Wang. Animationgpt:an aigc tool for generating game combat motion assets. https://github.com/fyyakaxyy/AnimationGPT, 2024. 1, 3

[45] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025. 2

[46] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *NeurIPS*, 2024. 2, 3, 4, 5

[47] Zeyu Ling, Bo Han, Yongkang Wong, Mohan Kangkanhalli, and Weidong Geng. Mcm: Multi-condition motion synthesis framework for multi-scenario. *arXiv preprint arXiv:2309.03031*, 2023. 2, 3, 6

[48] Zeyu Ling, Bo Han, Shiyang Li, Jikang Cheng, Hongdeng Shen, and Changqing Zou. Versatilemotion: A unified framework for motion synthesis and comprehension. *arXiv preprint arXiv:2411.17335*, 2024. 2, 3, 4, 5

[49] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *European conference on computer vision*, pages 612–630. Springer, 2022. 3

[50] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via masked audio gesture modeling. *arXiv e-prints*, pages arXiv–2401, 2023. 2, 3, 4, 6, 7

[51] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751, 2024. 3, 4

[52] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 4

[53] Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. Humantomato: Text-aligned whole-body motion generation. *ICML*, 2024. 3

[54] Shunlin Lu, Jingbo Wang, Zeyu Lu, Ling-Hao Chen, Wenxun Dai, Junting Dong, Zhiyang Dou, Bo Dai, and Ruimao Zhang. Scamo: Exploring the scaling law in autoregressive motion generation model. *arXiv preprint arXiv:2412.14559*, 2024. 2, 3, 4, 5

[55] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M^3GPT: An Advanced Multimodal, Multitask Framework for Motion Comprehension and Generation. *arXiv preprint arXiv:2405.16273*, 2024. 2, 3

[56] Jiageng Mao, Siheng Zhao, Siqi Song, Tianheng Shi, Junjie Ye, Mingtong Zhang, Haoran Geng, Jitendra Malik, Vitor Guizilini, and Yue Wang. Learning from massive human videos for universal humanoid pose control. *arXiv preprint arXiv:2412.14172*, 2024. 1

[57] Ian Mason, Sebastian Starke, and Taku Komura. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1):1–18, 2022. 3, 4

[58] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015: 18–24, 2015. 6

[59] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2, 4

[60] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 5

[61] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 2, 3, 4

[62] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *CVPR*, pages 722–731, 2021. 3, 5

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 8

[64] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 6, 8

[65] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *CVPR*, pages 11050–11059, 2022. 3

[66] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. In *ICLR*, 2022. 2, 3, 6, 7

[67] Shashank Tripathi, Omid Taheri, Christoph Lassner, Michael Black, Daniel Holden, and Carsten Stoll. Humos: Human motion model conditioned on body shape. In *European Conference on Computer Vision*, pages 133–152. Springer, 2024. 3

[68] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *CVPR*, pages 448–458, 2023. 3

[69] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: Probabilistic autoregressive dance generation with multimodal attention. *ACM Trans. Graph.*, 40 (6):195:1–195:14, 2021. 3

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 7

[71] Yabiao Wang, Shuo Wang, Jiangning Zhang, Ke Fan, Jiafu Wu, Zhucun Xue, and Yong Liu. Timotion: Temporal and interactive framework for efficient human-human motion generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7169–7178, 2025. 3

[72] Wikipedia contributors. Fbx. Wikipedia, The Free Encyclopedia, 2024. Last edited on 1 October 2024, at 12:38 (UTC). Accessed [insert access date]. 4

[73] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*, 2024. 3, 8

[74] Guowei Xu, Jiale Tao, Wen Li, and Lixin Duan. Learning semantic latent directions for accurate and controllable human motion prediction. In *European Conference on Computer Vision*, pages 56–73. Springer, 2024. 3

[75] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *CVPR*, pages 22260–22271, 2024. 2, 3, 4

[76] Kebing Xue, Hyewon Seo, Cedric Bobenrieth, and Guoliang Luo. Shape-conditioned human motion diffusion model with mesh representation. In *Computer Graphics Forum*, page e70065. Wiley Online Library, 2025. 3

[77] Hongdi Yang, Chengyang Li, Zhenxuan Wu, Gaozheng Li, Jingya Wang, Jingyi Yu, Zhuo Su, and Lan Xu. Smgdiff: Soccer motion generation using diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11807–11817, 2025. 3

[78] Han Yang, Kun Su, Yutong Zhang, Jiaben Chen, Kaizhi Qian, Gaowen Liu, and Chuang Gan. Unimumo: Unified text, music, and motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25615–25623, 2025. 3

[79] Payam Jome Yazdian, Eric Liu, Rachel Lagasse, Hamid Mohammadi, Li Cheng, and Angelica Lim. Motionscript: Natural language descriptions for expressive 3d human motions. *arXiv preprint arXiv:2312.12634*, 2023. 5

[80] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–480, 2023. 2, 3

[81] Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456, 2024. 3, 4

[82] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *CVPR*, pages 14730–14740, 2023. 3

[83] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 516–526, 2024. 3, 4

[84] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE TPAMI*, 2024. 3

[85] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. *arXiv preprint arXiv:2404.01284*, 2024. 2, 3, 4, 6

[86] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *AAAI*, pages 7368–7376, 2024. 3

[87] Zeyu Zhang, Yiran Wang, Wei Mao, Danning Li, Rui Zhao, Biao Wu, Zirui Song, Bohan Zhuang, Ian Reid, and Richard Hartley. Motion anything: Any to motion generation. *arXiv preprint arXiv:2503.06955*, 2025. 3

[88] Kaifeng Zhao, Gen Li, and Siyu Tang. Dart: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260*, 2024. 2, 3